3

Power Laws, Weblogs and Inequality by Clay Shirky

A persistent theme among people writing about the social aspects of weblogging is to note (and usually lament) the rise of an A-list, a small set of webloggers who account for a majority of the traffic in the weblog world. This complaint follows a common pattern we've seen with MUDs, BBSes, and online communities like Echo and the WELL. A new social system starts, and seems delightfully free of the elitism and cliquishness of the existing systems. Then, as the new system grows, problems of scale set in. Not everyone can participate in every conversation. Not everyone gets to be heard. Some core group seems more connected than the rest of us, and so on.

Prior to recent theoretical work on social networks, the usual explanations invoked individual behaviors: some members of the community had sold out, the spirit of the early days was being diluted by the newcomers, et cetera. We now know that these explanations are wrong, or at least beside the point. What matters is this: Diversity plus freedom of choice creates inequality, and the greater the diversity, the more extreme the inequality.

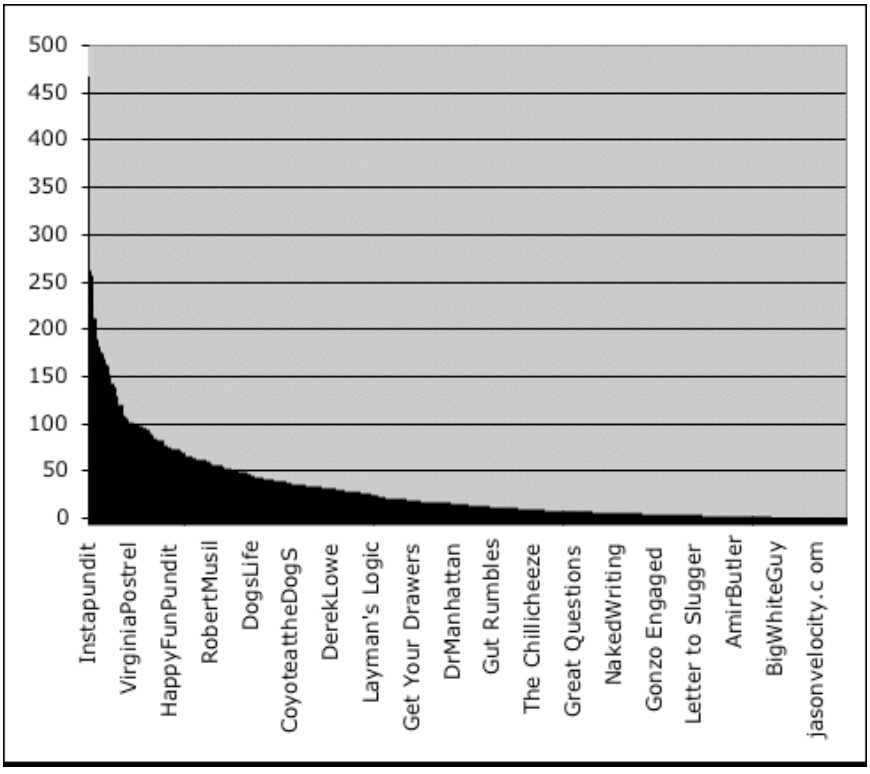
In systems where many people are free to choose between many options, a small subset of the whole will get a disproportionate amount of traffic (or attention, or income), even if no members of the system actively work towards such an outcome. This has nothing to do with moral weakness, selling out, or any other psychological explanation. The very act of choosing, spread widely enough and freely enough, creates a power law distribution.

A Predictable Imbalance

Power law distributions, the shape that has spawned a number of catch-phrases like the 80/20 Rule and the Winner-Take-All Society, are finally being understood clearly enough to be useful. For much of the last century, investigators have been finding power law distributions in human systems. The economist Vilfredo Pareto observed that wealth follows a "predictable imbalance", with 20% of the population holding 80% of the wealth. The linguist George Zipf observed that word frequency falls in a power law pattern, with a small number of high frequency words (I, of, the), a moderate number of common words (book, cat cup), and a huge number of low frequency words (peripatetic, hypognathous). Jacob Nielsen observed power law distributions in web site page views, and so on.

We are all so used to bell curve distributions that power law distributions can seem odd. The shape of Figure 1, several hundred blogs ranked by number of inbound links, is roughly a power law distribution. Of the 433 listed blogs, the top two sites accounted for fully 5 percent of the inbound links between them. (They were InstaPundit and Andrew Sullivan, unsurprisingly.) The top dozen (less than three percent of the total) accounted for 20 percent of the

Figure 1: 433 weblogs arranged in rank order by number of inbound links[1]



[1] The data is drawn from N.Z Bear's 2002 work on the blogosphere ecosystem. The current version of this project can now be found at http://www.myelin.co.nz/ecosystem/.

inbound links, and the top 50 blogs (not quite 12 percent) accounted for 50 percent of such links.

The inbound link data is just an example: power law distributions are ubiquitous. Yahoo Groups mailing lists ranked by subscribers is a power law distribution. (see Figure 2) LiveJournal users ranked by friends is a power law. (see Figure 3) Jason Kottke has graphed the power law distribution of Technorati link data. The traffic to this article will be a power law, with a tiny percentage of the sites sending most of the traffic. If you run a website with more than a couple dozen pages, pick any time period where the traffic amounted to at least 1,000 page views, and you will find that both the page views themselves and the traffic from the referring sites will follow power laws.

Figure 2: All mailing lists in the Yahoo Groups Television category, ranked by number of subscribers (Data from September 2002.)

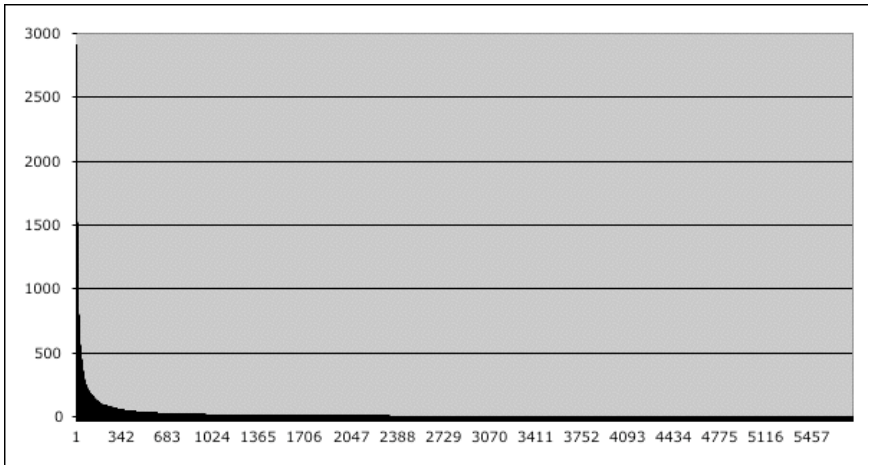
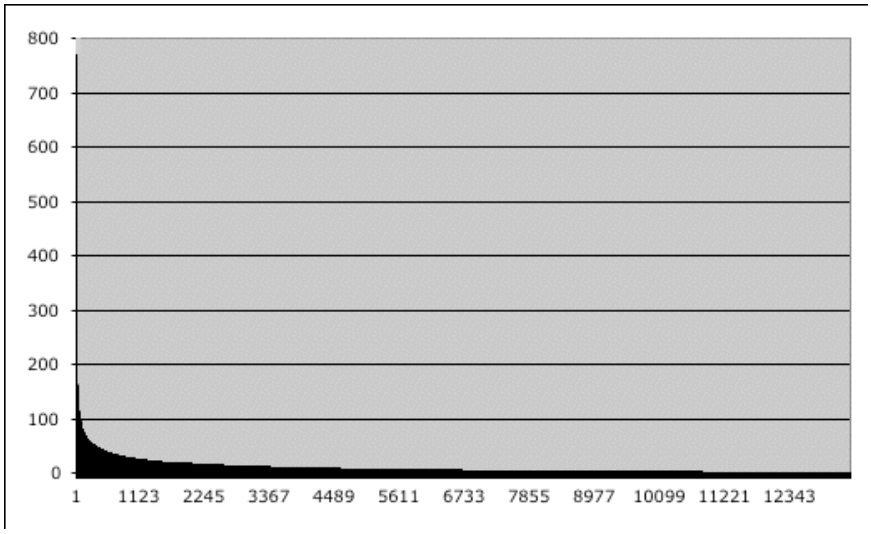


Figure #3: LiveJournal users ranked by number of friends listed. (Data from March 2002)



Rank Hath Its Privileges

The basic shape is simple: In any system sorted by rank, the value for the Nth position will be 1/N. For whatever is being ranked—income, links, traffic—the value of second place will be half that of first place, and tenth place will be one-tenth of first place. There are other, more complex formulae that make the slope more or less extreme, but they all relate to this curve. We've seen this shape in many systems. What've we've been lacking, until recently, is a theory to go with these observed patterns.

Now, thanks to a series of breakthroughs in network theory by researchers like Albert-Laszlo Barabasi, Duncan Watts, and Bernardo Huberman among others, breakthroughs being described in books like Linked, Six Degrees, and The Laws of the Web, we know that power law distributions tend to arise in social systems where many people express their preferences among many options. We also know that as the number of options rise, the curve becomes more extreme. This is a counter-intuitive finding; most of us would expect a rising number of choices to flatten the curve, but in fact, increasing the size of the system increases the gap between the Number One spot and the median spot.

A second counter-intuitive aspect of power laws is that most elements in a power law system are below average, because the curve is so heavily weighted towards the top performers. In Figure 1, the average number of inbound links (cumulative links divided by the number of blogs) is 31. The first blog below 31 links is 142nd on the list, meaning two-thirds of the listed blogs have a below-average number of inbound links. We are so used to the evenness of the bell curve, where the median position has the average value, that the idea of two-thirds of a population being below average sounds strange. (The actual median, 217th of 433, has only 15 inbound links.)

Freedom of Choice Makes Stars Inevitable

To see how freedom of choice could create such unequal distributions, consider a hypothetical population of a thousand people, each picking their 10 favorite blogs. One way to model such a system is simply to assume that each person has an equal chance of liking each blog. This distribution would be basically flat—most blogs will have the same number of people listing it as a favorite. A few blogs will be more popular than average and a few less, of course, but that will be statistical noise. The bulk of the blogs will be of average popularity, and the highs and lows will not be too far different from this average. In this model, neither the quality of the writing nor other people's choices have any effect; there are no shared tastes, no preferred genres, no effects from marketing or recommendations from friends.

But people's choices do affect one another. If we assume that any blog chosen by one user is more likely, by even a fractional amount, to be chosen by another user, the system changes dramatically. Alice, the first user, chooses her blogs unaffected by anyone else, but Bob has a slightly higher

chance of liking Alice's blogs than the others. When Bob is done, any blog that both he and Alice like has a higher chance of being picked by Carmen, and so on, with a small number of blogs becoming increasingly likely to be chosen in the future because they were chosen in the past.

Think of this positive feedback as a "preference premium." The system assumes that later users come into an environment shaped by earlier users; the thousand-and-first user will not be selecting blogs at random, but will rather be affected, even if unconsciously, by the preference premiums built up in the system previously.

Note that this model is absolutely mute as to why one blog might be preferred over another. Perhaps some writing is simply better than average (a preference for quality), perhaps people want the recommendations of others (a preference for marketing), perhaps there is value in reading the same blogs as your friends (a preference for "solidarity goods," things best enjoyed by a group). It could be all three, or some other effect entirely, and it could be different for different readers and different writers. What matters is that any tendency towards agreement in diverse and free systems, however small and for whatever reason, can create power law distributions.

Because it arises naturally, changing this distribution would mean forcing hundreds of thousands of bloggers to link to certain blogs and to de-link others, which would require both global oversight and the application of force. Reversing the star system would mean destroying the village in order to save it.

Inequality and Fairness

Given the ubiquity of power law distributions, asking whether there is inequality in the weblog world (or indeed almost any social system) is the wrong question, since the answer will always be yes. The question to ask is "Is the inequality fair?" Four things suggest that the current inequality is mostly fair.

The first, of course, is the freedom in the weblog world in general. It costs nothing to launch a weblog, and there is no vetting process, so the threshold for having a weblog is only infinitesimally larger than the threshold for getting online in the first place.

The second is that blogging is a daily activity. As beloved as Josh Marshall (TalkingPointsMemo.com) or Mark Pilgrim (DiveIntoMark.org) are, they would disappear if they stopped writing, or even cut back significantly. Blogs are not a good place to rest on your laurels.

Third, the stars exist not because of some cliquish preference for one another, but because of the preference of hundreds of others pointing to them. Their popularity is a result of the kind of distributed approval it would be hard to fake.

Finally, there is no real A-list, because there is no discontinuity. Though explanations of power laws (including the ones here) often focus on numbers like "12 percent of blogs account for 50 percent of the links", these are arbitrary markers. The largest step function in a power law is between the Number One and Number Two positions, by definition. There is no A-list that is qualitatively different from their nearest neighbors, so any line separating more and less trafficked blogs is arbitrary.

The Median Cannot Hold

However, though the inequality is mostly fair now, the system is still young. Once a power law distribution exists, it can take on a certain amount of homeostasis, the tendency of a system to retain its form even against external pressures. Is the weblog world such a system? Are there people who are as talented or deserving as the current stars, but who are not getting anything like the traffic? Doubtless. Will this problem get worse in the future? Yes.

Though there are more new bloggers and more new readers every day, most of the new readers are adding to the traffic of the top few blogs, while most new blogs are getting below average traffic, a gap that will grow as the weblog world does. It's not impossible to launch a good new blog and become widely read, but it's harder than it was last year, and it will be harder still next year. At some point (probably one we've already passed), weblog technology will be seen as a platform for so many forms of publishing, filtering, aggregation, and syndication that blogging will stop referring to any particularly coherent activity. The term "blog" will fall into the middle distance, as "home page" and "portal" have, words that used to mean some concrete thing, but which were stretched by use past the point of meaning. This will happen when head and tail of the power law distribution become so different that we can't think of J. Random Blogger and Glenn Reynolds of Instapundit as doing the same thing.

At the head will be webloggers who join the mainstream media (a phrase which seems to mean "media we've gotten used to.") The transformation here is simple - as a blogger's audience grows large, more people read her work than she can possibly read, she can't link to everyone who wants her attention, and she can't answer all her incoming mail or follow up to the comments on her site. The result of these pressures is that she becomes a broadcast outlet, distributing material without participating in conversations about it.

Meanwhile, the long tail of weblogs with few readers will become conversational. In a world where most bloggers get below average traffic, audience size can't be the only metric for success. LiveJournal had this figured out years ago, by assuming that people would be writing for their friends, rather than some impersonal audience. Publishing an essay and having three random people read it is a recipe for disappointment, but publishing an account of your Saturday night and having your three closest

friends read it feels like a conversation, especially if they follow up with their own accounts. LiveJournal has an edge on most other blogging platforms because it can keep far better track of friend and group relationships, but the rise of general blog tools like Trackback may enable this conversational mode for most blogs.

In between blogs-as-mainstream-media and blogs-as-dinner-conversation will be thought of as "Blogging Classic," blogs published by one or a few people, for a moderately-sized audience, with whom the authors have a relatively engaged relationship. Because of the continuing growth of the weblog world, more blogs in the future will follow this pattern than today. However, these blogs will be in the minority for both traffic (dwarfed by the mainstream media blogs) and overall number of blogs (outnumbered by the conversational blogs.)

Inequality occurs in large and unconstrained social systems for the same reasons stop-and-go traffic occurs on busy roads, not because it is anyone's goal, but because it is a reliable property that emerges from the normal functioning of the system. The relatively egalitarian distribution of readers in the early years had nothing to do with the nature of weblogs or webloggers. There just weren't enough blogs to have really unequal distributions. Now there are.

Addendum: *David Sifry, creator of the Technorati.com, has created the Technorati Interesting Newcomers List, in part spurred by this article. The list is designed to flag people with low overall link numbers, but who have done something to merit a sharp increase in links, as a way of making the system more dynamic.*